10/6/95

AN 1996:69350 CAPLUS
DN 124:136911
TI SEARCH: Locating and identifying genes and intron-exon junctions in large genomic sequences using parallel computers
AU Jorgensen, Joshua; Buskirk, Stewart F.; Gribskov, Michael; Smith, Douglas W.
CS Department Philosophy, University California, La Jolla, CA, 92093, USA
SO Bioinf. Genome Res., Proc. Int. Conf., 3rd (1995), Meeting Date 1994, 145-59. Editor(s): Lim, Hwa A.; Cantor, Charles R. Publisher: World Scientific, Singapore, Singapore.
CODEN: 62FLAC
DT Conference
LA English
AB The purpose of the SEARCH program is to locate the position and identify the function of genes in large DNA sequences via comparison with protein sequences in protein libraries using parallel computers such as the Intel iPSC/860 and Paragon. The SEARCH algorithm is a two-step windowing algorithm in which 25-mer "windows" in a lookup table prepd. from the six reading frames of the query sequence are compared with 25-mers from library sequences, first requiring an identity of 2 residues (ktup of 2) and then using the PAM250 distance matrix for scoring. Parallel processors are used by assigning different library entry sequences to different processors, with no message passing between processors. Timing expts. indicate high scalability, with a near linear response to 64 nodes on the iPSC/860, and a speed approx. 45 times that of the Cray YM/P at 64 nodes. The ability of SEARCH to find and functionally identify genes has been tested using the phage lambda genome (48 kb), the ECOUW85U E. coli 95 kb sequence, and yeast chromosome III (315 kb). Sensitivity and specificity tests have been performed by analyzing the ability of SEARCH to find functionally similar proteins using the Prosite ***database*** defined set of protein families and by comparison of sequences from the globin superfamily. Similar globin tests were performed with BLASTX, for comparative purposes. Tests of the ability of SEARCH to identify intron-exon junctions in eukaryotic DNA sequences have been made using globin sequences. A graphic-based interface for Macintosh computers, with anal. capability for interpretation of SEARCH results, has been developed. The windowing approach used by SEARCH provides a fast and robust software tool for comparison of large genomic sequences to protein libraries, and for locating and identifying features in sequence and other character-based ***databases*** in general, on both parallel and nonparallel platforms.